JOURNAL OF CULTURAL ANALYSIS AND SOCIAL CHANGE

Journal of Cultural Analysis and Social Change,
2025, 10(4), 5028-5053
ISSN: 2589-1316

Is Data Really the New Oil?

Szilárd Malatyinszki[1*], Bálint Kocsis[2], Szilvia Módosné Szalai[3], Botond Géza Kálmán[4]

[1] Habil, Ph.D, Associate Professor, Department of Management and Business Law, Faculty of Economics, John von Neumann University, doctoral supervisor, John von Neumann University Doctoral School of Management and Business Administration, 10 Izsáki út, Kecskemét HU-BA-6000, HUNGARY, e-mail: malatyinszki.szilard@nje.hu, https://orcid.org/0000-0002-1624-4902, e-mail: malatyinszki.szilard@gmail.com

[2] Student, Kodolányi János University, Faculty of Economics

[3] Assistant Professor, Széchenyi István University, Kautz Gyula Faculty of Economics, 9026, Győr, Hungary

[4] Assistant Professor of the Department of Finance and Accounting (DFA) in the Faculty of Economics of John von Neumann University (NJE GTK), address: 10 Izsáki út, Kecskemét HU-BA-6000, HUNGARY, e-mail: kalman.botond.geza@nje.hu;

[4] Visiting Researcher of the Department of Accounting and Auditing in the Ferenc Rakoczi II Transcarpathian Hungarian College of Higher Education // Zakarpats'kij ugors'kij institut imeni Ferenca Rakoci II (KMF), address: 6 Ploshcha Kossuth, Berehove UA-90201, Transcarpathia, UKRAINE, e-mail: kalman.botond@kmf.org.ua; https://orcid.org/0000-0001-8031-8016, PhD, e-mail: botondgezakalman@gmail.com

*__Corresponding Author:__ malatyinszki.szilard@gmail.com

Citation: Syaifudin, A., Hendarmawan, Sunardi, Novianti, E. & Hariadi, H. (2025). Is Data Really the New Oil?, *Journal of Cultural Analysis and Social Change*, 10(4), 5028-5053. https://doi.org/10.64753/jcasc.v10i4.4189

Published: December 29, 2025

ABSTRACT

This study examines whether data can truly be considered "the new oil" by analysing the strategic role of data assets and data-driven decision-making in modern organisations. The research explores how raw data can be transformed into valuable insights through data analytics and machine learning, and how these insights contribute to improved corporate performance, competitiveness, and strategic decision-making. The theoretical framework is based on data asset management, data science, and the CRISP-DM methodology, which provides a structured approach to extracting business value from data. The empirical part of the study presents a case study from the telecommunications sector, focusing on customer churn prediction. Using publicly available data, the research applies logistic regression and random forest models to demonstrate how analytical tools can predict customer behaviour and support targeted retention strategies. The results show that advanced machine learning methods significantly outperform traditional models in identifying customers at risk of churn, highlighting the tangible economic value of data-driven approaches through cost reduction and revenue stabilisation. Beyond the firm level, the study also examines the relationship between digitalisation and national competitiveness by comparing Hungary, Estonia, and Romania using international indicators such as DESI, DII, and GDP. The findings reveal that while digitalisation contributes to competitiveness, the relationship is complex and non-linear, influenced by institutional, human capital, and structural factors. Overall, the research concludes that data only becomes a valuable resource when it is properly collected, processed, analysed, and embedded into decision-making processes. Organisations and national economies that invest in digital capabilities, analytical expertise, and data culture gain a sustainable competitive advantage in the digital age.

Keywords: Data assets, data-driven decision-making, machine learning, customer churn prediction, digital competitiveness

INTRODUCTION

Data and its strategic use are one of the cornerstones of the modern business world. For companies, the data they collect is not only a source of information, but also a valuable resource that predicts market trends and customer preferences and facilitates data-driven decision-making. Over the past few decades, the concept of data assets and their strategic application, as well as data-driven decision-making processes, have received increasing attention in many economic and political organisations around the world. The paradigm shift in business and social decision-making is moving towards data-driven approaches, which are essential in today's rapidly changing world. Due to the embeddedness of digital devices and platforms in our everyday lives, we leave behind a trail of data points. We walk around with our mobile phones constantly connected to the network in our pockets, we pay in shops with our bank cards, we consume content with our browsers linked to our Google accounts, or we search the internet on Instagram, which is linked to our Facebook accounts. We like holiday photos or the latest watches, coffee makers or cars. Our online activities inevitably leave digital footprints behind us. The financial sector, commerce, industry and transport also generate data points during their operations. This could be the route and fuel consumption data of a truck in a transport company's fleet, the operating temperature and noise level of various components of a gas turbine, the transaction logs of a bank's customers' accounts, or the list of products entered into a restaurant chain's cash register during the day. (Orlando Troisi, 2019)

"The purpose of data analytics is to use analysis to generate meaningful information from the raw data at our disposal, which we can use to support our strategic and operational decisions, draw conclusions, and thus help optimise the performance of our organisation." (Davenport, 2014)

This research examines the application of data analytics and machine learning using the example of publicly available data from a telecommunications company. Using the methodology presented, the aim of the research is to show how valuable insights can be gained from the available data to facilitate decision-making and improve corporate performance, which can mean, for example, higher turnover, greater customer satisfaction, or more effective advertising campaigns.

The English term "insight" is used, which better conveys the understanding of the deeper connections inherent in the data. By analysing the role of data-driven decision-making in the strategic use of data, such as who to target with promotions, the research contributes to understanding and addressing the business challenges of the digital age .

A well-structured data analysis process, such as that provided by *the "CRoss-Industry Standard Process for Data Mining, hereinafter CRISP-DM"* (IBM, 2016) framework, enables companies to maximise the information and deeper understanding derived from data. This can mean, for example, more accurate forecasts, the identification of patterns and groups that can be targeted more effectively, or the emergence of time series trends that have not been taken into account in decision-making until now.

I will illustrate subscriber churn prediction using the example of an anonymous telecommunications company and show how data can be used to generate value, such as revenue.

The strategic use of data is key to business success, as it enables companies to respond proactively to change and improve their decision-making processes. For example, if an international agricultural machinery distributor starts collecting failure data in all countries and then compares this data with the climatic conditions at the place of use, a failure map can be drawn up showing which parts are more prone to failure in colder conditions. It can then consider equipping its machines in those regions with other parts that are more resistant to cold. Data-driven decision-making reduces costs by enabling targeted changes and can increase user satisfaction by reducing breakdowns.

"*Like oil, data must be processed, 'refined', interpreted, and the information gained from the data must be used appropriately in order for our data to become valuable and our organisation to become more successful. Collecting, storing, cleaning, processing, and finally presenting and interpreting information are all resource-intensive tasks.*" (Forbes, 2022) However, digitalisation and the data-driven decision-making that can be built on it can be an effective tool not only for companies but also for other economic actors, such as national economies. It can increase competitiveness and GDP, reduce the bureaucratic burden on the population, and shorten administrative procedures.

In the literature review section of my research, I examine the state of digitalisation in the public and private sectors in Hungary. I compare our competitiveness with that of other countries in the region, the European Union and the OECD, with a more detailed comparison of Romania and Estonia. I examine the countries' gross domestic product (GDP), digital intensity level (DII) and Digital Economy and Society Index (DESI). (European Commission, 2023)

Before researching the above topics, I seek answers to the following hypotheses:

Analytical tools can indeed be used to generate meaningful, commercially useful information from the appropriate raw data. In this case, I examine whether the data available to a telecommunications company can be

used to produce reliable user churn predictions using logistic regression and random forest models, which can help with targeted customer retention campaigns.

With the information generated from the available raw data using analytical tools, the future can be predicted under certain circumstances and to a certain extent. For example, when performing the above analysis, how confidently can we determine that a given customer will churn?

In countries where digitalisation and data-driven decision-making are more advanced and society is more knowledge-based according to the DESI and DII indices, competitiveness, such as GDP growth, is higher.

In my research, I will use theoretical materials, experience gained at my workplace dealing with data analysis, and two examples of my own data analysis to confirm or refute the first two hypotheses. In the data analysis, I examined several aspects, or attributes, such as the development of domestic and international per-minute charges, the costs of night and daytime calls, and the number of minutes spoken. In my research, I also discuss the most common data types and their characteristics. I will confirm or refute the third hypothesis using available international statistics.

Problem Statement and Motivation

My own experiences influenced the problem statement. I sought answers to questions that I encounter on a daily basis. How can the available data be utilised, and why do companies often fail to do so? Why is digitalisation important, even if it means an additional burden for a business or even a country? These questions provided the basis for me to delve deeper into the topic and examine such a situation from a practical point of view.

Goals and Usefulness

I structured my research in such a way as to illustrate the challenges, present a tangible solution to the problem through data analysis, and place the multidisciplinary nature of data asset utilisation in a broader perspective with an international outlook. This is an area where specialised knowledge, such as programming, is not enough; many different fields of expertise must be involved throughout the entire process in order to maximise the benefits, which may include increased revenue, cost reduction, market acquisition or retention, for example. Choosing an inappropriate analysis model can result in a customer retention campaign targeting more people than necessary, thus increasing the cost of the campaign. Economists, business analysts, data miners and industry experts can all contribute to a data-driven decision-making process. Due to its scope limitations, the research does not aim to teach programming skills or comprehensive analytical methodology.

Tasks

In order to achieve the set goal, the following tasks need to be completed.

- Review of the literature
- Selecting the methodology required for the research
- Selection of source data
- Selecting two models for analysis
- Preparation of data analysis
- Selection of countries used for international comparison
- Narrowing down the available statistics to the topic under investigation

Structure of the Research

In terms of logical structure, I sought to first provide a general overview and define what data assets are and how the concept developed. This helps us to better understand the challenges and risks associated with data assets. After a theoretical presentation of the CRISP framework, I guide the reader through a possible practical application in the form of a case study. The data analysis of the telecommunications company described earlier continues the thread in demonstrating the effectiveness of data-driven decision-making. In describing the modelling I have carried out, I also refer to the relevant literature for easier understanding. Leaving the microeconomic environment behind, the international perspective helps me to examine and demonstrate the impact of digitalisation at the national economic level as well. The examples of Hungary, Romania and Estonia illustrate the complexity and not necessarily linear relationship between digitalisation and competitiveness. Finally, I make practical recommendations for cooperation on this topic, which may be useful for professionals working in the competitive sector and in education. Due to the subject matter of the research, there are many English expressions that are not always translated in the profession, and the original English is used, or has already become part of everyday language. Wherever possible, I will try to provide an explanation and translation for the expression.

Target Groups

Each part of the research could be a separate field of research, but in many cases I have only touched on a topic because I wanted to provide the reader with a broader perspective and, at the end, make suggestions that I believe could be useful and thought-provoking for all professionals who work in a similar field or would like to learn more about this area.

LITERATURE REVIEW

The interpretations available in the literature help to understand the theoretical background and key concepts of the research. In the following chapters, I will discuss the topics that cover the terms used in the subsequent analyses. Their development also plays an important role in understanding them, so I will also provide a historical overview.

The Importance and Challenges of Data Assets

The concept of data assets has become increasingly important with the development of information technology and the digital economy. Treating data as a strategic resource means that companies can derive value and competitive advantage from it. *"Analysing and utilising data allows companies to gain deeper insights into customer preferences, market trends and internal operational efficiency. In this context, data assets refer not only to the totality of raw data, but also to the information that can be extracted from this data."* (Davenport, 2014) Economic actors should view this as a resource in the same way as human capital or the components used in manufacturing. The roots of data and information management date back to the early days of commercial activities and financial records. Manual accounting systems and statistics were already in use in ancient times and the Middle Ages, essentially involving the use of data to support management and financial decision-making. With the technological advances of the 20th century, particularly the spread of computing and the internet, the possibilities for collecting, storing and analysing data have grown exponentially. The concept of data assets began to emerge during this period, when companies realised that data was not only used to record operational activities, but also represented a valuable resource. The connection with financial data is particularly strong, as financial performance analysis, return on investment calculation, budget planning and financial risk management are fundamentally data-driven processes. Analysing financial data enables companies to make more accurate forecasts, improve the quality of financial decision-making and optimise their resources. However, with the advance of digitalisation, data is being generated in more and more places, whether from sensors, transaction logs, software usage or purchasing habits. There are few areas of life left where we do not collect some form of digital information.

Data assets play a prominent role in competitiveness. Companies are able to integrate insights from data into strategic decision-making, anticipate market changes, and deliver personalised customer experiences. In addition, it enables the development of new business models and product innovation. According to a report by the McKinsey Global Institute, data-driven organisations achieve greater profitability and market value than those that do not exploit the potential of data. (Bughin, 2017) Through digital transformation, organisations are collecting ever-increasing amounts of data, which represents enormous value if properly managed and analysed. However, this requires the right professionals, because the mere existence of large amounts of data is only a starting point; without processing and interpretation, it has almost no value. In our fast-paced world, the speed at which we can process this data has also become an important consideration. When a financial report appears on the managing director's phone, how quickly can a stockbroker access an economic analysis?

For example, the New York Stock Exchange was the first to recognise that the speed of data analysis plays a key role, directly influencing trading strategies, decision-making and market positions. Trading decisions are often made within seconds or even milliseconds, so rapid data analysis is essential to gain a competitive advantage. This is one of the best examples of algorithmic trading, where the speed of data analysis plays a critical role . Algorithms perform trading operations based on predefined rules, responding quickly to changes in market data. The speed of data analysis allows traders to immediately take advantage of market fluctuations before they have a significant impact on prices. (ERIK, 2011) High-frequency trading (HFT) is another area where the speed of data analysis is crucial. HFT companies use computer algorithms to execute a large number of trading operations in a very short time, often in milliseconds or microseconds. The speed of data analysis allows them to gain an advantage by exploiting market trends and price changes.

In general, data can provide a competitive advantage through the following:

- Better decision-making: Based on data, organisations can make more accurate forecasts, identify trends and allocate resources more efficiently. Analysing past data can help identify mistakes and success factors, thereby improving the quality of future decisions.

- Becoming more innovative: It can be used to develop new products and services, optimise existing ones and increase efficiency. Data-driven innovation can help organisations open up new market niches, stay ahead of competitors and adapt flexibly to changing market demands.
- More personalised customer experience: Organisations can better understand customer needs, preferences and behaviour. This enables them to provide personalised offers, services and communications, improving customer satisfaction and loyalty. My research will provide deeper insights into the practical application of customer retention and related forecasting.
- Stronger market position: Data-driven decision-making and innovation can help organisations strengthen their market position, enter new markets and stay ahead of competitors by enabling them to make faster, more reliable, fact-based decisions. (McKinsey & Company, 2023)

Quantifying the above benefits could, of course, be a separate area of research, as measuring them can mean different things for different businesses.

Risks Associated with Data Assets

The complexity of data collection and storage stems from the fact that companies must be able to quickly and efficiently record, store and retrieve large amounts of data. This includes structured data such as customer names and transaction information, as well as unstructured data such as social media posts or videos. Choosing the right data storage solutions, including on-premises data centres and cloud-based storage, requires significant IT infrastructure and expertise and can be costly. In addition, data management and the rules governing it must also be taken into account. In the European Union, the GDPR (General Data Protection Regulation) governs the protection of natural persons' data. Therefore, any collection and transfer of such data is subject to this legislation, which entails additional expert and legal costs.

Data security means protecting data assets from unauthorised access, use, destruction and cyber security threats. As the amount of digital data increases, so does the risk of cyber attacks, data leaks and data theft. Companies need to implement advanced security protocols and technologies such as encryption, firewalls and access management mechanisms to protect their data. Even large companies can fall victim to cyber attacks despite their precautions, which often exploit the carelessness of unsuspecting employees to gain access to critical systems. In terms of data security, ransomware attacks currently pose one of the greatest risks, especially in 2023, where these attacks continue to pose a significant threat to companies worldwide. (SentinelOne Blog, 2023) Ransomware, or blackmail viruses, are malicious software that lock the victim's data or systems and demand a ransom for restoring the data or unlocking the system. Such an attack can even bring production plants or factories to their knees. In my work, I have seen examples of such an attack bringing an international manufacturing company with a modern IT infrastructure to its knees. As a result of the attack, several factories were shut down, causing huge losses for the company. In addition to ransomware, other threats such as phishing scams, zero-day exploits, fileless malware and Denial-of-Service (DoS) attacks also pose a significant challenge to the security of corporate systems. These attacks exploit vulnerabilities in computers, smartphones and internet-connected devices, further increasing the cybersecurity challenges for companies.

Data Science, Machine Learning and the CRIPS Framework

"Data science is a multidisciplinary field that uses scientific methods, procedures, algorithms and systems to gain insights and knowledge from structured and unstructured data. It combines expertise from various fields, such as statistics, mathematics, computer science and domain-specific knowledge, to analyse and interpret complex data sets." (James, 2023)

Statistical learning or machine learning is a subset of the aforementioned data science. Statistical learning encompasses a wide range of tools for understanding data, classified into supervised and unsupervised categories. In supervised statistical learning, the emphasis is on creating a statistical model that predicts or estimates an output (target variable, dependent variable) based on one or more inputs (explanatory variables, independent variables). Problems of this type arise in various fields, such as business, medicine, astrophysics and public policy. In contrast, unsupervised statistical learning deals with inputs that do not have supervised outputs (no target variable), but allows the discovery of relationships and structures within the data. (James, 2023)

The CRIPS-DM (CRoss-Industry Standard Process for Data Mining) framework helps us understand the process by which data science, and especially statistical learning, is integrated into corporate operations today. The following *(Figure 1)* illustrates the different steps of this process.

- Business understanding: Defining business objectives, assessing the current situation, defining data mining objectives and preparing a project plan.
- Data understanding: Collecting initial data, preparing data descriptions, exploring data samples, and checking data quality.

- Data preparation (often the most time-consuming stage): Selecting relevant data, cleaning and structuring data, integrating data sets, and formatting data appropriately.
- Modelling: Selecting a modelling technique, designing tests, building the model and evaluating its performance.
- Evaluation: Analysing the results, reviewing the entire process, and determining the next steps.
- Deployment: Developing deployment plans, outlining monitoring and maintenance strategies, preparing a final report, and reviewing the project.

In the rest of my research, I will apply this framework to analyse the case study example. First, I will present the task, then perform exploratory data analysis, generate new variables, build different models, evaluate them, and finally report on the potential integration of the models.

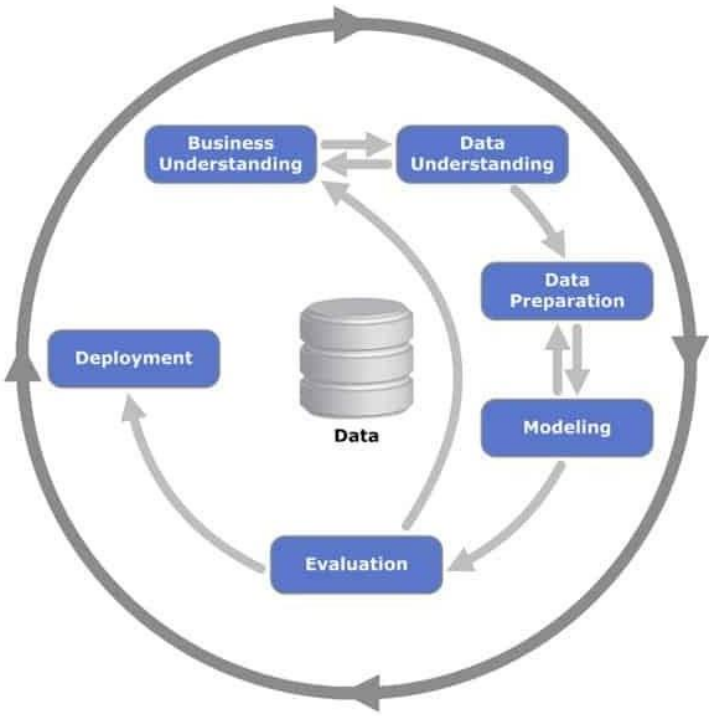


Figure 1 CRISP – Source: Datascience-pm.com

Translation: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, Deployment

The Emergence of Machine Learning

The history of machine learning dates back to the 1950s, when Arthur Samuel defined it as an early form of artificial intelligence. Over the years, the development of machine learning has grown exponentially, particularly due to the availability of data, increased computing capacity and the development of algorithms. Applications of machine learning are now widespread, from healthcare and finance to manufacturing and services. (IBT, 2023) Arthur Samuel, an IBM engineer, first used the term "machine learning" when he taught an IBM 701 computer to play draughts. Alan Turing then published his famous article, "Computing Machinery and Intelligence," in which he posed the question, "Can machines think?" and introduced the famous Turing test. The purpose of the test was to test a machine's ability to exhibit human-like intelligent behaviour. The test was not based on whether the machine gave the correct answers to the questions, but on how similar its answers were to human answers. (Stanford, 2021)

Machine learning enables systems to learn from data, recognise patterns and make decisions with minimal human intervention. In the 1960s and 70s, there was growing interest in neural networks and models that attempted to model the functioning of the human brain. Frank Rosenblatt developed an early type of neural network. Despite early successes, however, research into neural networks declined in the 1980s, partly due to limitations in data and computing capacity.

The 1990s brought significant advances in data mining and statistical modelling. Parallel to the development of data storage and processing technology, the applications of machine learning also expanded. Decision trees, support vector machines (SVM), and clustering algorithms opened up new possibilities in data analysis. (Coenen, 2021)

Since the early 2010s, deep learning, a specific type of machine learning based on deep neural networks, has revolutionised the field. Deep learning has made it possible to tackle problems that previously seemed unsolvable, such as image recognition, natural language processing, and improving performance in complex games. Google Brain, DeepMind, and other research groups have ushered in a new era in machine learning with their breakthroughs. (Pichai, 2023)

Today, machine learning applications are widespread in almost all areas of artificial intelligence. However, ethical and privacy issues, algorithmic bias and the black box problem pose new challenges for researchers.

Most Common Data Types

Understanding the properties of data is essential for case studies, as modelling itself can be built on these foundations. Data types are categories into which data is classified. Each data type has its own characteristics, such as the range of values that can be stored, the operations that can be performed on them, and the amount of memory they occupy. Programmers need to be aware of the characteristics of a given data type in order to program efficiently and reliably. In programming, data types play a key role in the storage, processing, and analysis of data. The data type determines what values the data can store and what operations can be performed on it. Below is a brief overview of the most common data types, illustrated with examples. (Luenendonk, 2022)

Numeric Data Types:

- Integers: Store whole values without decimal places. For example: 1, 2, 3, -10.
- Floating point numbers: Store real numbers that include decimal places. For example: 3.14, -2.718.

Text Data Types:

- Character strings: Store letters, numbers, and symbols. For example: "apple", "12345", "!@#$%^&*".

Logical Data Types:

- True/False: Stores Yes/No or On/Off values. Date and time types:
- Dates: Store dates. For example: 2 December 2023.
- Times: Store times. For example: 10:29:00. Categorical data types:
- Categories: Store data classified into different categories. For example: "male", "female", "red", "blue", "green".

Binary Data Types:

- Binary data: Stores 0s and 1s. Arrays or sequences that store data in binary format, such as images or files.

The most common data can be classified into two main categories: structured and unstructured data. Structured data has a well-defined format and is easy to search, such as tabular data stored in databases. Unstructured data is less organised, such as text documents, images or videos. In addition, there is semi-structured data, such as JSON or XML files, which are structured but more flexible. Quantitative and qualitative data types, which are based on numbers or descriptive characteristics, are also often used in data analysis. (Nelson, 2020)

Case Study Description

For my research, I chose to predict subscriber churn as a case study. To do this, I downloaded a sample database from kaggle.com containing data from a telecommunications company, specifically which subscribers are likely to switch providers.

Customer churn, i.e. the termination of customer service subscriptions, is a critical metric for assessing customer satisfaction and the overall health of the company. The churn rate, calculated as the percentage of customers lost within a specified period, is a key indicator that stands in stark contrast to the customer growth rate, which tracks the number of new customers. (Datacamp, 2023)

Beyond natural and seasonal churn, which refers to normal business operations, various factors can indicate underlying problems within the company: (Glimsdahl, ICMI, 2021)

- Inadequate or poor customer support: Insufficient or inadequate customer support can lead to dissatisfaction and churn.
- Negative customer experiences: Unfavourable interactions or experiences may prompt customers to cancel their subscriptions.
- Switching to a competitor: It is a threat if customers choose a competitor that offers better terms or prices.
- Changing priorities: Customers customers' priorities change may can lead to the discontinuation of services.
- Long-term dissatisfied customers: Even loyal customers may leave if they are not satisfied over time.
- Unmet expectations: If the service does not meet customer expectations, they may leave.
- Financial problems: Economic constraints or problems can contribute to churn.
- Fraud protection for payments: Overly strict fraud protection measures related to customer payments can lead to dissatisfaction.

The high customer churn rate significant challenges significant a company:

- Revenue loss: Churn is directly related to revenue loss, which affects the financial stability of the company.
- Acquisition costs: Acquiring new customers is more expensive than retaining existing ones, especially in highly competitive markets.
- Damage to reputation: Poor customer service leading to churn can damage a company's reputation through negative reviews on social media or review websites.

Customer retention is becoming a key element of business strategy, especially for subscription services. Analysing customer behaviour data, including purchase intervals, lifetime value, cancellations, post-purchase interactions and online activity, is essential for predicting and managing customer churn. Identifying the characteristics of at-risk customers enables proactive measures to be taken. Machine learning, particularly classification models, can help predict churn . (Datacamp, 2023)

In this section, I have presented the business relevance of predicting customer churn. In the next section, I will present the data I am using as a case study. I performed the data analysis using the Python language and the Jupyter Notebook environment. The figures were generated in the Jupyter Notebook environment, which limits the number of labels, so I have included the axis names in the figure titles.

Data Description

The dataset contains cleaned customer activity characteristics and a churn label that indicates whether the customer has churned. This real-world scenario demonstrates the practical usefulness of applying data-driven approaches to reduce customer churn and increase business sustainability.

In this research, I refer to the data source downloaded from kaggle.com as (Diamantaras, 2020).

Descriptive Statistics

The data from kaggle.com consists of 4,250 rows and 20 columns. The rows represent customers, i.e. the objects under study, while the columns represent the data generated by customers, or in other words, the attributes associated with each customer.

Description of columns: name in English as it appears in the database, explanation and unit of measurement.

Table 1 Column description, source: kaggle.com

Serial number	Column name	Description	Unit of measurement
1	state	US Abbreviation of the state	text value
2	account_length	how long the user has been with the provider	number of months
3	area_code	3-digit area code	number
4	international_plan	Yes/No	logical
5	voice_mail_plan	Yes/No	logical
6	number_vmail_messages	Number of voicemail messages	number
7	total_day_minutes	Total minutes of calls during the day	minutes
8	total_day_calls	Number of daytime calls	number

9	total_day_charge	Charge for daytime calls	dollars
10	total_eve_minutes	Evening call time in minutes	minutes
11	total_eve_calls	Number of evening calls	number
12	total_eve_charge	Evening call charges	dollars
13	total_night_minutes	Night call minutes	minutes
14	total_night_calls	Number of night calls	number
15	total_night_charge	Charge for night calls	dollars
16	total_intl_minutes	International call time in minutes	minutes
17	total_intl_calls	Number of international calls	number
18	total_intl_charge	Charge for international calls	dollars
19	number_customer_service_calls	Number of calls to customer service	per
20	churn	Yes/No	logical

Using the pandas Python package, we can view the first five rows of data. Here, I am displaying the current data transposed for better clarity. (Figure 2.)

	0	1	2	3	4
state	OH	NJ	OH	OK	MA
account_length	107	137	84	75	121
area_code	area_code_415	area_code_415	area_code_408	area_code_415	area_code_510
international_plan	no	no	yes	yes	no
voice_mail_plan	yes	no	no	no	yes
number_vmail_messages	26	0	0	0	24
total_day_minutes	161.6	243.4	299.4	166.7	218.2
total_day_calls	123	114	71	113	88
total_day_charge	27.47	41.38	50.9	28.34	37.09
total_eve_minutes	195.5	121.2	61.9	148.3	348.5
total_eve_calls	103	110	88	122	108
total_eve_charge	16.62	10.3	5.26	12.61	29.62
total_night_minutes	254.4	162.6	196.9	186.9	212.6
total_night_calls	103	104	89	121	118
total_night_charge	11.45	7.32	8.86	8.41	9.57
total_intl_minutes	13.7	12.2	6.6	10.1	7.5
total_intl_calls	3	5	7	3	7
total_intl_charge	3.7	3.29	1.78	2.73	2.03
number_customer_service_calls	1	0	2	3	3
churn	no	no	no	no	no

Figure 2 First 5 rows of data, units of measurement according to the previous table

We can see that most variables are either float or int, and there are a total of four columns whose values can be categorical. These are state, area_code, international_plan, voicemail_plan, and our target variable, churn. To validate our assumption, let's see

How many unique values belong to the different columns.

Table 2 Number of unique values, source: own

Serial number	Column name	Value	Unit of measurement
1	'account_length'	215	pieces month
2	'international_plan'	2	logical
3	'voice_mail_plan'	2	logical
4	'number_vmail_messages'	46	pieces
5	'total_day_minutes'	1843	minutes
6	'total_day_calls'	120	pieces
7	'total_day_charge'	1843	dollars
8	'total_eve_minutes'	1773	minutes
9	'total_eve_calls'	123	pieces
10	'total_eve_charge'	1572	dollars
11	'total_night_minutes'	1757	minutes
12	'total_night_calls'	128	pieces
13	'total_night_charge'	992	dollars
14	'total_intl_minutes'	168	minutes
15	'total_intl_calls'	21	calls
16	'total_intl_charge'	168	dollars
17	'number_customer_service_calls'	10	pieces
18	'churn'	2	logical

In the following, I will analyse categorical and continuous variables separately and examine how they relate to the target variable in both cases. We can call these attributes.

First, let us examine the target variable. The values for churn, i.e. which subscribers switched telephone services, are illustrated in the following figure. Here, the proportion of users who switched providers is 14% (meaning that 86% of users remained with the same provider). We call this an unbalanced distribution. A balanced distribution would mean a 50-50% ratio in this case. (Datacamp, 2023)

In the following figures, the Y axis (Count) shows the number of objects, i.e. users.

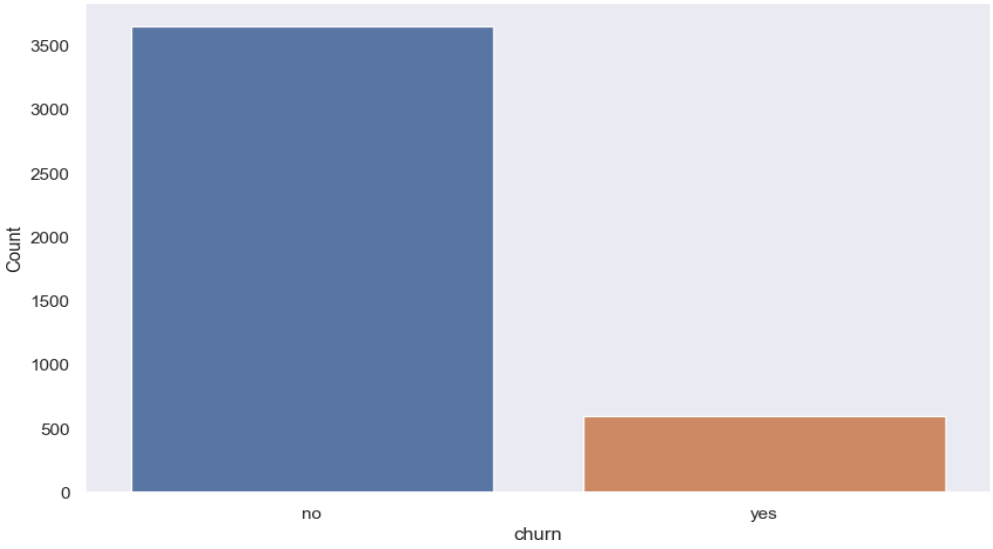


Figure 3 Distribution of the Churn variable. X-axis: churn yes/no, y-axis: number of users (own figure)

The distribution of the area_code variable is illustrated in the following figure. The proportion of switching users is 14% in area 408, 13.6% in area 415, and finally 15% in area 510.

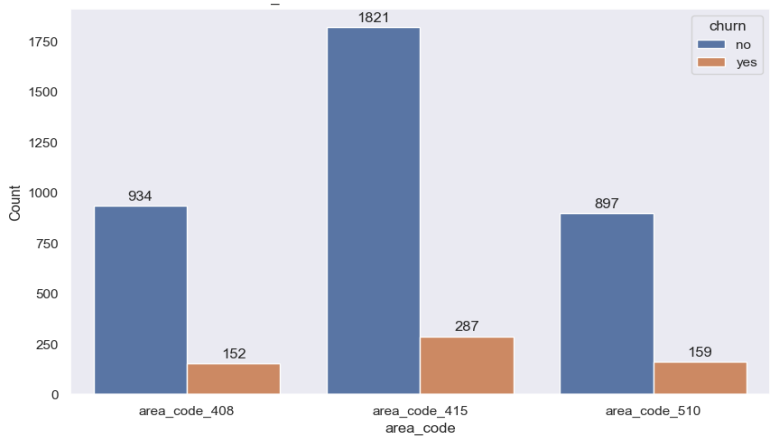


Figure 4 Joint distribution of the area_code and churn variables. X-axis: churn yes/no, area code; Y-axis: number of users (own figure)

In the case of the international_plan variable, the proportion of switching users is 42% among those who have this service and 11% among those who do not. Here we can see that the international_plan variable probably plays an important role in predicting churn, which may be due to cheaper services offered by competitors.

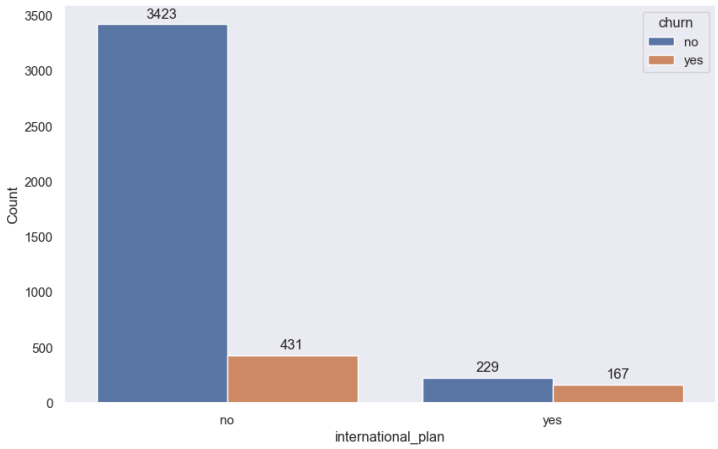


Figure 5 Joint distribution of the international_plan variable and churn (own figure)

In the case of the voice_mail_plan variable, the churn rate among users who have a subscription is 16%, while among those who do not have one, it is 7% of the total number of subscribers.

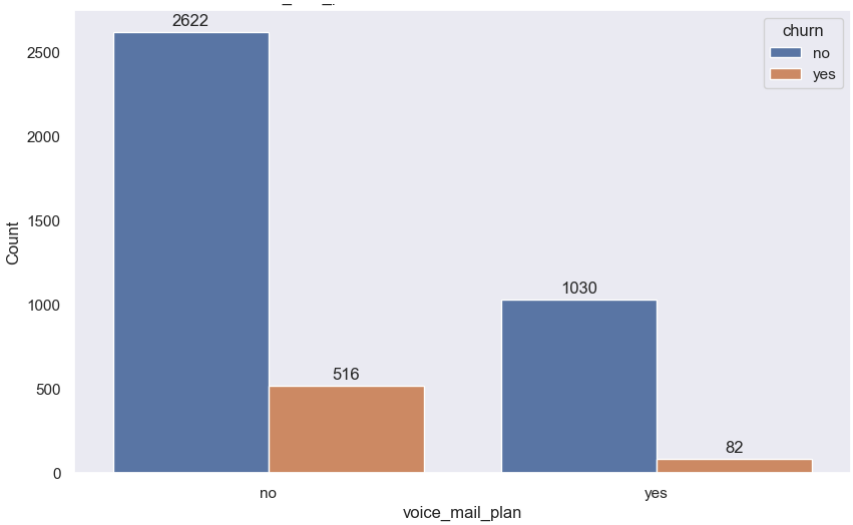


Figure 6 Joint distribution of the voice_mail_plan variable and the churn variable (own figure)

The distribution of the state variable shows the number of churners for each state. The figure shows, for example, that West Virginia has the highest number of subscribers, with 120 people who did not switch providers and 19 who did. Based on the figure, the churn rate is highest in NJ (New Jersey) (27%).

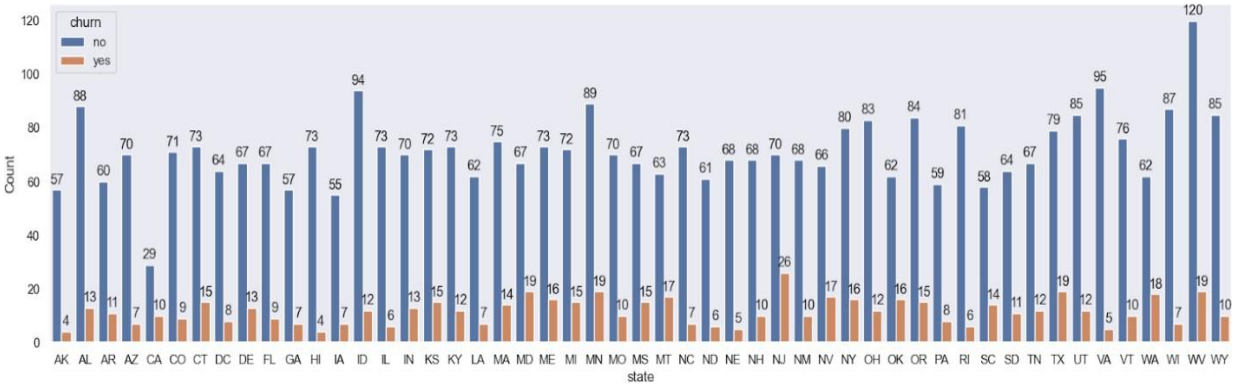


Figure 7 Joint distribution of the state variable and the churn variable; (own figure)

In the case of continuous variables, I would like to highlight just a few. There is a clear relationship between the total_day_minutes and total_day_charge variables: the more someone phones, the higher the cost.

The average fee among users who switch providers is 35.5, while among those who do not switch providers it is 29.8 (shown in blue in the figures), which represents a difference of approximately 15%. This value is 17.8 for total_eve_charge among users who switched providers and 16.8 among those who did not switch. For the total_intl_charge variable, this value is 2.87 among users who churned and 2.75 among those who did not churn.

The distributions for the variables presented are illustrated in the following figures. In the figures, the X-axis shows the quantity expressed in the unit of measurement for the attribute, while the Y-axis shows the distribution. Blue indicates non-churn users, while red indicates churn users.

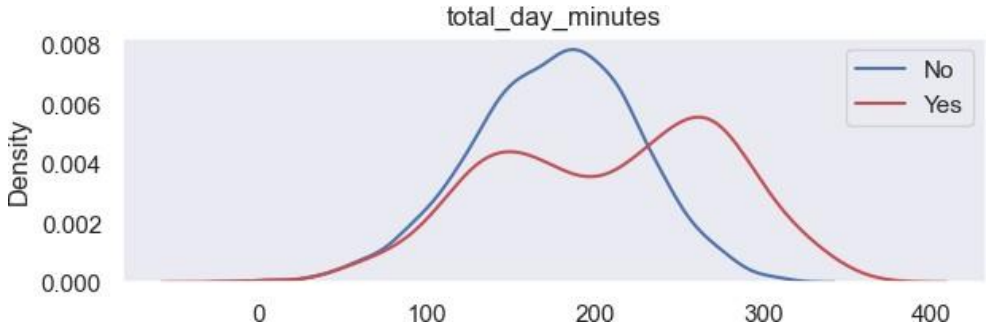


Figure 8 Duration of daytime calls in minutes (own figure)

Figure 8 shows that non-churn and churn users follow different distribution curves in terms of the length of daytime calls. This may indicate different user habits, which may provide an opportunity for segmentation.

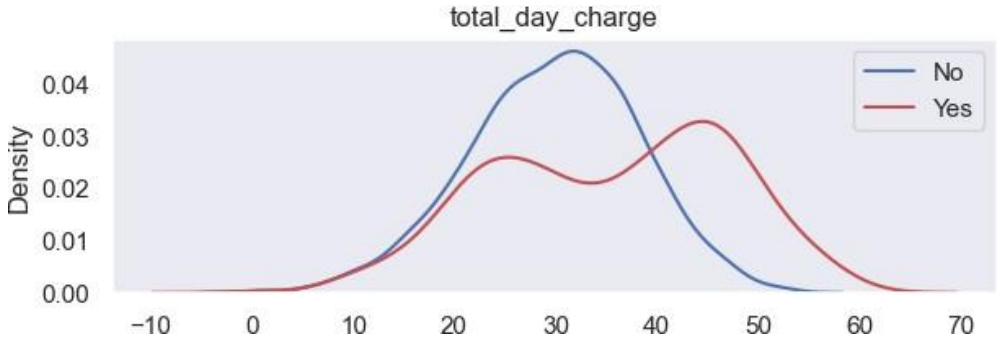


Figure 9 Distribution of daytime call charges in dollars (own figure)

Figure 9 shows a similar difference. From this, we can conclude that the models will be able to make good use of these factors to produce more accurate predictions.

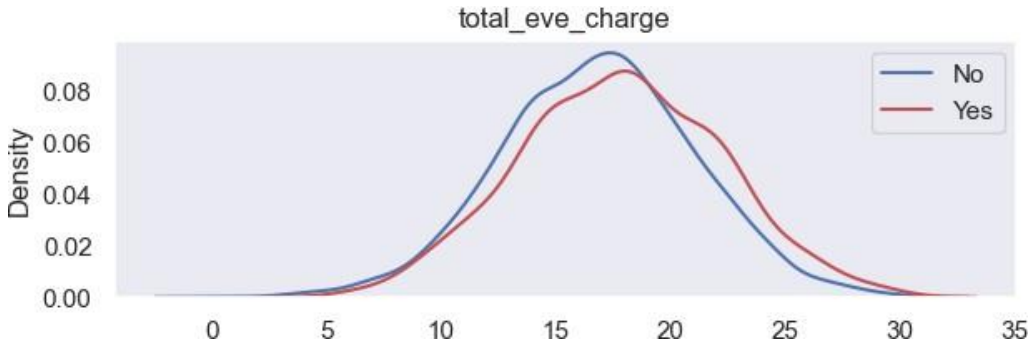


Figure 10 Distribution of evening call times in minutes (own figure)

Figures 10 and 11 show that there is minimal deviation in the distributions of the examined attributes compared to the other two attributes. In these cases, it is less likely to separate the objects that are characterised by attrition.

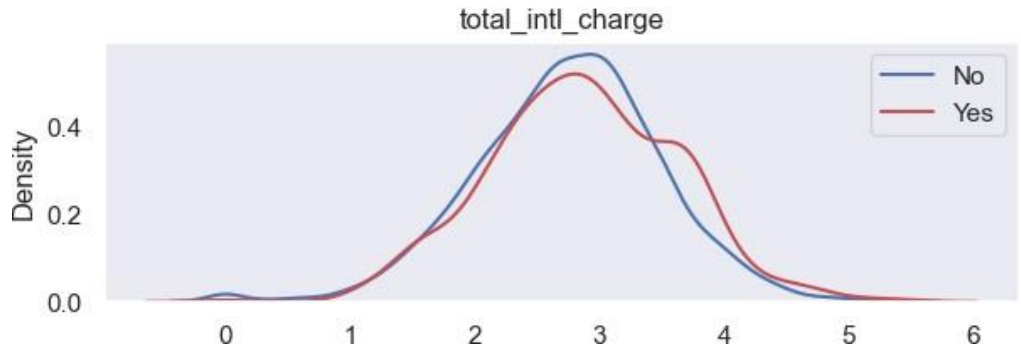


Figure 11 Distribution of international call charges in dollars (own figure)

Data Preparation

Before modelling, I performed the following data transformations. I recoded categorical variables that only take two values (international_plan, voice_mail_plan, churn) to values 0 and 1. For categorical variables that take on multiple values (state, area_code), I used one-hot encoding. This means that the values that appear in a column are placed in a separate column using binary encoding based on the values they take in the original column. This is illustrated in the following figure.

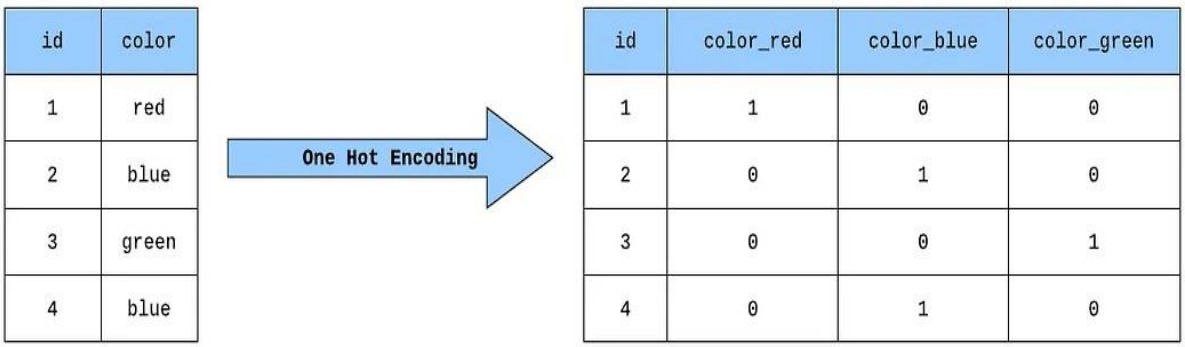


Figure 12 Variable encoding (Novack, 2020)

For continuous variables, I generated the following variables. I created the variable total_net_minutes, which is the sum of total_day_minutes, total_eve_minutes, total_night_minutes and total_intl_minutes. This is essentially the sum of all variables. In addition, I created the total_net_calls variable (the sum of total_day_calls, total_eve_calls, total_night_calls, and total_intl_calls) and the total_net_charge variable (the sum of total_day_charge, total_eve_charge, total_night_charge, and total_intl_charge) in a similar manner. The distribution of the variable according to churn is illustrated by the following variable. The averages are 62 for churn and 55 for non-churn users. Here, too, we can see that there is a 10% difference between the averages of the two groups.

Emberek	Előrejelzés	Valós érték
1	P	P
2	P	P
3	N	N
4	N	N
5	N	N
6	N	P
7	P	N
8	P	N
9	P	N
10	P	N

Konfúziós Mátrix	Vizsgált eredmény (előrejelzés/ predikció) Pozitív (P)	Vizsgált eredmény (előrejelzés/ predikció) Negatív (N)
Valós állapot Pozitív (P)	2	1
Valós állapot Negatív (N)	4	3

Figure 13 Distribution of total costs (own figure)

Classification Evaluation and the Aim of the Competition

The aim of the competition between the models is to predict as accurately as possible which users will switch providers, and for this we will use the accuracy metric, which can be calculated as follows. Accuracy can be defined using a confusion matrix, which is commonly used to evaluate the classification predictions of machine learning algorithms. (Provost, 2013) The confusion matrix has two axes. One is what we predicted on a binary output scale (based on an example from medicine, whether a given patient is sick or not according to our model), and the other is the patient's actual condition (positive or negative, i.e., whether they are sick or not in real life). (Wikipedia, 2023) Based on this, we can evaluate our predictions (the results of our machine learning model) according to four categories

- True negative: the patient is not sick and is not sick according to our prediction.
- True positive: the patient is ill and our prediction also indicates that they are ill.
- False positive: someone is sick based on our prediction, but is not sick in reality. (For example, we say that a man is pregnant.)
- False negative: based on our prediction, someone is not ill, but in reality they are (for example, someone's COVID test came back negative, but in reality they are ill and infected, so the test was wrong, or we predict that an 8-month pregnant woman is not pregnant).

Accuracy can be obtained as the ratio of the sum of correctly predicted events to the number of elements in the sample, i.e.:

Accuracy = sum(true negatives, true positives)/ sum(true negatives, true positives, false negatives, false positives)

		vizsgálat eredménye		
		negatív (–)	pozitív (+)	
valós állapot (arany standard)	egészséges (–)	valós negatív (VN)	álpozitív (ÁP)	→ specificitás VN/(VN+ÁP)
	beteg (+)	álnegatív (ÁN)	valós pozitív (VP)	→ szenzitivitás VP/(ÁN+VP)
		↓ szegregancia VN/(VN+ÁN)	↓ relevancia VP/(ÁP+VP)	pontosság

Figure 14 Accuracy and precision analysis (Wikipedia, 2023)

For example, if we test 10 people for COVID-19, for whom we gave the following predictions (tested result), and the actual values (actual status) are as follows, based on our confusion matrix, the accuracy is 0.5, which we obtain based on ((2+3) / 10)

Figure 15 Confusion Matrix (Wikipedia, 2023)

Modelling and Results

In this chapter, I present the results of the modelling I wrote and performed myself. I used two different models, one was a logistic regression and the other was a random forest. To evaluate the model, I used a 20% training-testing sample split, which means that I trained the model on 80% of the available data and evaluated its performance on the remaining 20%. In the following, I present the theoretical background of the two models and the results achieved.

Logistic Regression

"Logistic regression is a statistical method used for binary classification, which means that it predicts the probability that an observation belongs to one of two possible outcomes. Despite its name, logistic regression is more of a classification than a regression algorithm. It models the relationship between a dependent binary variable and one or more independent variables by estimating the probability of an event occurring." (James et al., 2023)

The logistic regression model uses the logistic function, also known as the sigmoid function, to convert a linear combination of input features into a value between 0 and 1. This output represents the probability of the observation belonging to the positive class. The model is trained by adjusting the weights of the input features through an optimisation process, typically using the maximum likelihood estimation method.

Logistic regression is widely used in various fields, including finance, healthcare, and marketing, as it is simple, interpretable, and effective for handling binary classification tasks. It serves as one of the basic algorithms of machine learning and is often chosen when the relationship between input features and binary outputs needs to be explored and used for prediction. ( )

The data asset on which the analysis is based consists of the aforementioned 4,250 rows and 20 columns. The rows represent the customers, i.e. the objects under investigation, and the columns represent the data generated by the customers, or in other words, the attributes associated with the customer. In this case, we are examining 20% of this, which is 850. Figure 16 shows how the model breaks down this user quantity in the matrix. We can evaluate the individual fields according to the previous Figure 15.

The following confusion matrix contains the results of the logistic regression. We can see that out of the 129 churn users in the test sample, our model only found 13. Although the accuracy is 0.845, this value is distorted by the fact that the model was able to correctly predict the majority of users who did not switch providers. However, by examining sensitivity, we can get a more accurate picture of our model's performance. This shows how many of the users who churned we were able to find. The value of this is 13 / (13+116), is 0.1.

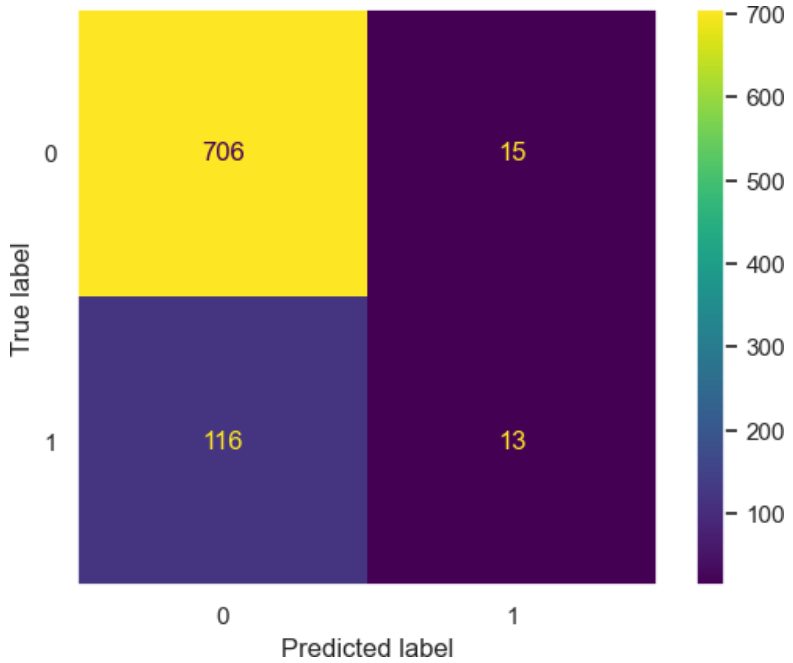


Figure 16 Logistic regression results. The colour scale represents the number of users according to the scale on the right (own figure)

Table 3 Logistic regression results

True positive	706	15
True negative	116	13
	Prediction positive	Prediction negative

Table 3 shows the row and column explanations for the previous figure and the corresponding values representing the number of users.

Next, I'll show you a random forest classification algorithm and compare it with the results of logistic regression.

Random Forest

"The Random Forest classifier is a versatile and effective ensemble learning algorithm that is widely used in machine learning for both classification and regression tasks. It belongs to the family of decision tree-based models. The "forest" in its name refers to a collection or ensemble of decision trees, while "random" refers to the incorporation of randomness in both the training and prediction processes." (James, 2023)

In a random forest, multiple decision trees are trained on different subsets of the data set, and each tree makes its own independent predictions. During training, random subsets of features are also considered for each tree, adding an additional layer of variability. This randomness helps prevent overfitting and contributes to the robustness of the model. To make predictions, Random Forest combines the outputs of individual trees by majority voting (for classification) or averaging (for regression). This ensemble approach often results in better generalisation performance and higher accuracy compared to individual decision trees. Random Forest is known for its ability to handle high-dimensional data sets, capture complex relationships, and automatically handle feature selection. They are less prone to overfitting and can handle noisy or irrelevant features. This makes random forests a popular choice for various applications, such as finance, healthcare, and image recognition, where accurate and reliable predictions are crucial. (James, 2023)

The following figure illustrates the confusion matrix for random forests. This figure also shows 20% of the entire database, i.e. 850 users. The accuracy is 0.979, which is much better than the result of logistic regression. We can also see in the figure that our tree-based model finds many more users who switch providers. The sensitivity value is also much higher, at 0.86. Based on this, it can be clearly concluded that random forest achieves significantly better results than logistic regression on the present data.

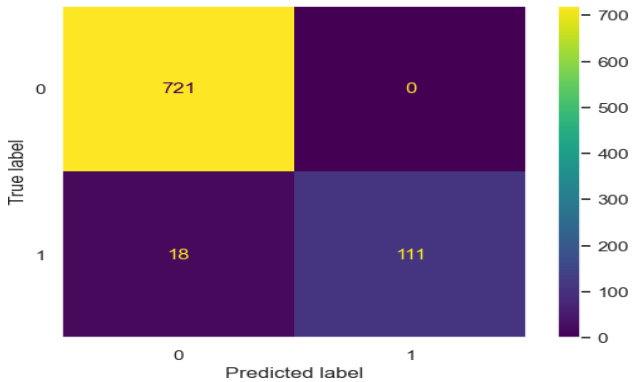


Figure 17 Random Forest Results The colour scale represents the number of users according to the scale on the right (own figure)

The results can be interpreted as shown in the figure below

Konfúziós Mátrix	Vizsgált eredmény (előrejelzés/ predikció) Pozitív (P)	Vizsgált eredmény (előrejelzés/ predikció) Negatív (N)
Valós állapot Pozitív (P)	721	0
Valós állapot Negatív (N)	18	111

Figure 18 Random Forest confusion matrix with the number of users examined (own figure)

RESULTS

In my research, I chose to predict user churn in the telecommunications sector as a case study. First, I presented the CRIPS framework, which I used to process the data. Then I presented the task and its business relevance, more specifically, why it is worthwhile for telecommunications companies to predict which users will switch providers. I examined the data based on various descriptive statistics and identified a few trends that may help to understand why users switch providers. I performed data transformations and presented the evaluation method for machine learning classification prediction, and predicted the target variable using two different models. The results of the modelling highlighted that it is not necessarily worthwhile to choose accuracy as the evaluation metric. The reason for this is the distribution of the target variable, namely , as only 15% of the users in the sample switched providers, which represents an unbalanced sample. In this sample, random forest proved to be the better model.

In further use of the case study, for example, it is easy to imagine that if we can accurately predict which users will switch providers, we can even prevent this with various discounts. For example, as we saw in the descriptive statistics, the introduction of fee discounts can even have a positive effect on user retention.

Let us examine a hypothetical simplified situation according to the table below.

	Number of users	Coupon	Coupon cost
Everyone	10,000	10	$ 100,000.00
Predicted 15%	1,500	$ 1	$ 15,000.00
		Difference:	$ 85,000

We have 10,000 users, and we are launching a customer retention campaign because we are seeing that they are all churning. We don't have good data analysis, so we send every user a £10 voucher to improve satisfaction. This will cost us £100,000. However, if we had been able to perform data analysis similar to the above, we could have narrowed this campaign down to the 15% who are most likely to churn. That's only 1,500 people. This would save the company £85,000.

It is clear, therefore, what a competitive advantage it can be for a company to be able to exploit its own data assets and make decisions accordingly. This is a clear example of monetisation, because if we can target people who are highly likely to churn with special offers, we can save costs. The more targeted our messages are, the greater the expected results. However, in order to perform such analyses, we need high-quality and up-to-date data. To achieve this, a company must reach a high level of digitalisation. It must invest in the right enterprise management system, which can be flexibly tailored to emerging data needs. It is necessary to develop a well-functioning business intelligence framework where data is quickly accessible on a daily basis and is filtered and cleaned according to appropriate logic. In order to create value from data, reports must be relevant and provide decision-makers with adequate insight. In other words, data should not only provide us with mere information, but also with in-depth understanding or insight that helps us uncover, predict and interpret the correlations and trends behind phenomena, thus supporting informed decision-making.

Developing and maintaining such a system also requires the right expertise. IT specialists who are capable of building a modern and scalable framework. Business analysts who have industry knowledge and can translate customer needs into business logic, create data models and analyse the relationships between data sources. Data scientists may be needed to create predictive models and deeper analyses to make even better use of data assets.

These are all areas that are closely related to the digitalisation and competitiveness of a given region or country. The better a country's digitalisation status, the greater the chance that a company will find such professionals and that the infrastructure will be sufficiently developed so that it does not suffer a disadvantage in international terms. A culture of digitalisation must also be present at the company level, because those who fail to develop in this area will fall behind in the market.

The relationship between digitalisation and competitiveness

Digitalisation is transforming all areas of the economy. The beneficial effects of digitalisation can be observed at the macroeconomic level and even in the smallest businesses. However, knowledge of the level of digitalisation alone is not sufficient to assess the competitiveness of countries. The digital performance of European Union member states and the development of European digital competitiveness are measured by the DESI (Digital Economy and Society Index). (European Commission, 2023) This is a fairly complex indicator that takes into account several areas of digital development:

- Connectivity: Measures the availability and quality of broadband infrastructure, including mobile broadband connections.

- Human capital: Examines the level of digital skills and competences, from basic user skills to software development.
- Internet usage: Measures various aspects of the population's online activities, including online communication, use of social networks, online shopping and content consumption.
- Digital technology integration: Measures the extent to which digital technologies such as e-commerce, cloud services and big data are used in the business sector.
- Digital public services: Assesses the availability and quality of e-government and e-health services, including demand for digital services.

Based on research, understanding the relationship between the DESI index and gross domestic product (GDP) in the EU is complex. According to one study, there is a weak but positive linear relationship between the DESI index, which assesses the digital progress of EU Member States, and average GDP growth. Based on this, relative digital development may in some cases lead to faster economic growth, although this relationship has not always proven to be statistically significant. (Vyshnevskyi, 2021) According to the study, the digitisation of the EU's digital economy has not proven to be a decisive driver of economic growth, suggesting that although digitisation is important, it does not in itself guarantee high economic growth rates.

Comparison of Countries

The following example illustrates the complexity of the issue. Based on DESI data available on the European Commission's website, I have selected three countries with different levels of digitalisation. I examine the situation in Hungary, Estonia and Romania in the category of Integration of Digital Technologies. (European Commission, 2023)

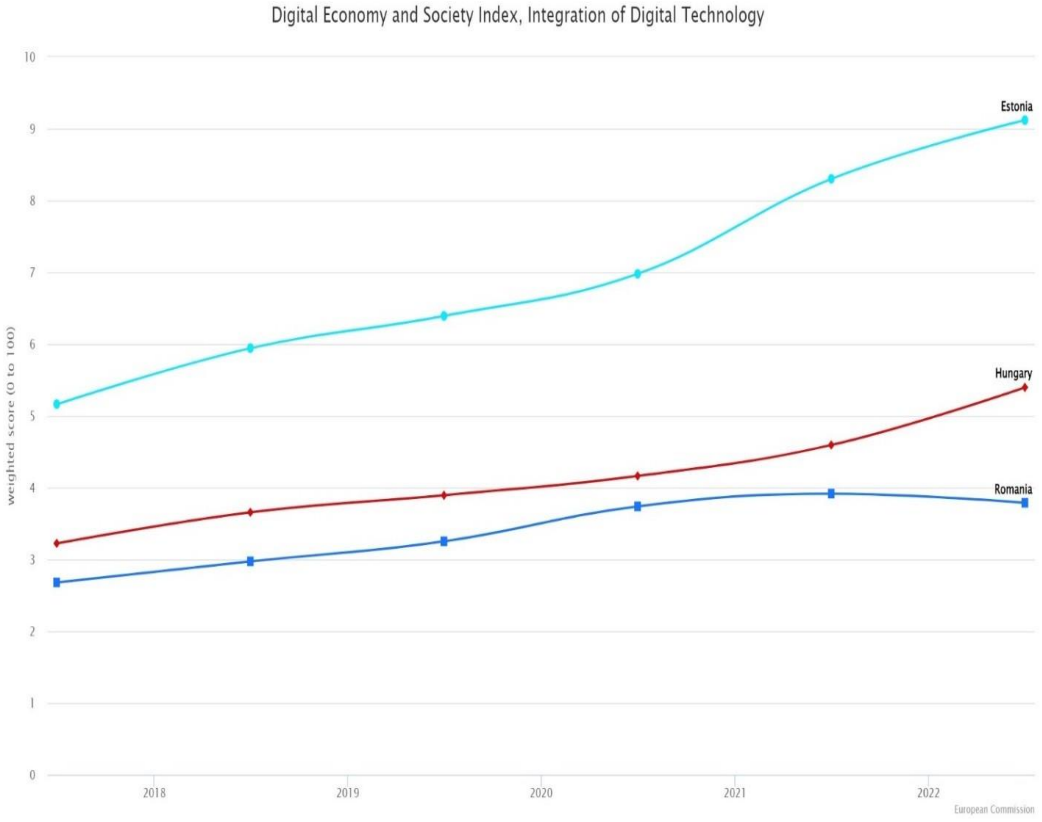


Figure 19 DESI; Integration of digital technology, X: years, Y: DESI index Source: (European Commission, 2023)

The figure above clearly shows that Estonia is well ahead in the digitalisation race. Hungary is ahead of Romania.

The following figure shows the GDP of the same countries for the same period.

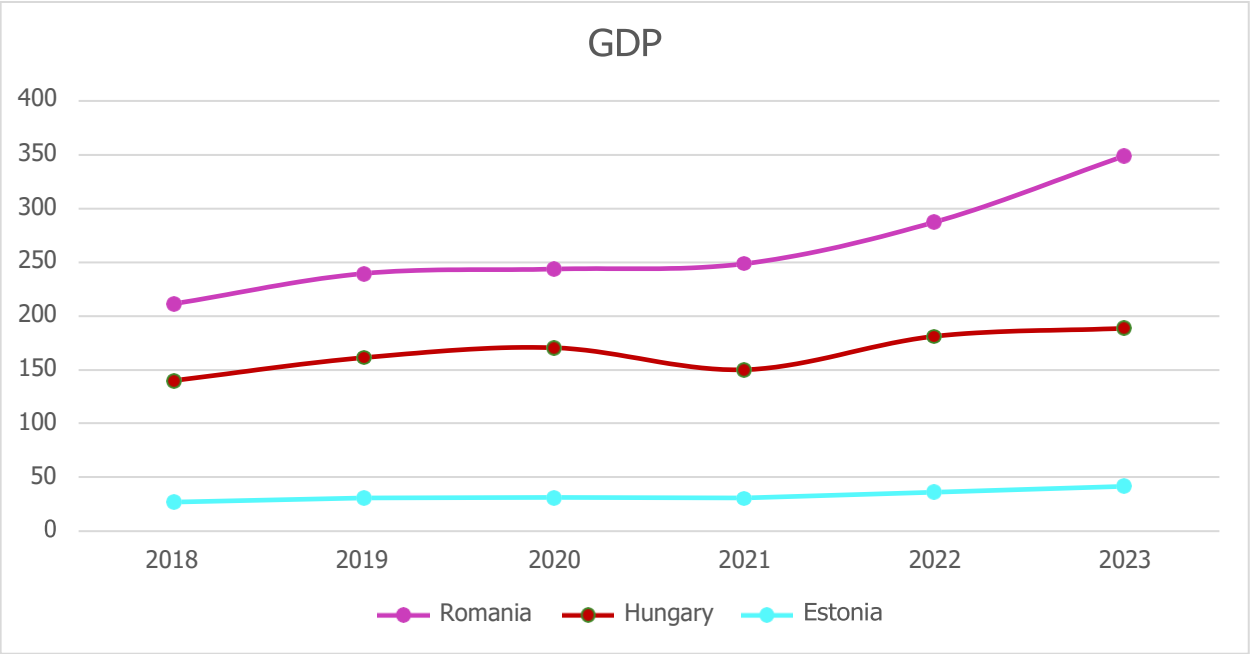


Figure 20 GDP of selected countries, X: years, Y: GDP (Eurostat, 2023)

Here, we can see a reversal of the order. Estonia's GDP lags behind that of the other two countries. Of course, at first glance, we could say that the size of the countries and their different economic conditions provide an obvious explanation for this, but this does not explain one thing, which is the trend. Both Hungary and Estonia have progressed faster and at a higher rate in their digital transformation than Romania. Despite this, neither country has been able to fully exploit this advantage in terms of GDP growth. However, GDP alone does not necessarily reflect a country's competitiveness and prosperity. It is therefore worth examining each country in more detail.

Estonia

Estonia is at the forefront of the digital economy and has even been nicknamed the "digital tiger". Behind the success of digitalisation lies a well-thought-out strategy, early investment and the active involvement of citizens. The key to success lay in early investment. In the 1990s, the Estonian government invested heavily in building internet infrastructure, making the internet accessible to its citizens. The aim of the "Tiger Leap" programme was to bring computers to schools and the general public, laying the foundations for the widespread dissemination of digital knowledge. The country boasts the highest internet usage rate in the EU, with e-services being widely available and 99% of the population conducting their business online. In addition, Estonia has become a leading player in the digital economy, home to a number of innovative start-ups. (Szentmihályi, 2023)

Here are a few examples of how an excellent digital environment can help companies get started. Estonia is home to world-famous companies such as Skype, TransferWise (Wise) and Bolt, which have revolutionised communication, financial services and urban transport. Although Skype's founders, Niklas Zennström and Janus Friis, are of Danish and Swedish origin, the software itself is the work of Estonian talent. Ahti Heinla, Jaan Tallinn and Niklas Zennström developed the programme in 2003, and its success soon took off. Skype was acquired by Microsoft in 2011 for $8.5 billion, but the company's development centre remained in Estonia, employing thousands of people in Tallinn and Tartu. Estonia is not only home to Skype. TransferWise (Wise) has revolutionised the online money transfer market with its low fees and fast transactions, while Bolt (Taxify) dominates the online taxi booking market in Eastern Europe and Africa. Estonia's example shows that even a small country can become a leader in the digital economy if it applies a well-thought-out strategy, commitment and a citizen-centric approach. The story of the "digital tiger" can be an inspiration to other countries around the world that are on the path to becoming digital societies.

Estonia's economic competitiveness and GDP growth are characterised by innovation, digital transformation and openness. The country excels in the integration of digital technologies, which contributes to its high level of economic competitiveness. Estonia is a leader in e-government and e-administration services, which improve citizen access and increase efficiency. GDP growth is stable, reflecting the strong foundations of the economy and growth driven by innovative businesses. The country's economic model focuses on continuous development and the application of new technologies.

Romania

Romania has made significant progress towards a digital economy in recent years, but there is still much to be done to realise its full potential and bridge the digital divide. Despite an early internet boom, economic recession and a lack of investment slowed development in the 1990s. In the 2000s, the government recognised the potential of the digital economy and began investing in IT infrastructure and the development of e-services. The 2010s focused on the digital economy, with the development of strategic plans and the expansion of internet access for the population. More and more people began to use e-services, and the start-up ecosystem also began to develop. Currently, Romania is below the EU average in terms of digital economy development. (European Commission, 2023) The digital divide remains significant in rural areas and among lower-income groups. The government continues to invest in digital infrastructure and the development of e-services, with the aim of bringing Romania to the forefront of the European digital economy by 2030. The key to success may lie in developing digital skills, increasing investment and bridging the digital divide. However, broadband internet coverage is currently below the EU average, making it difficult for people to take better advantage of digital opportunities and creating obstacles to the development of businesses. Nevertheless, progress and convergence are already evident.

As part of my work, I also deal with the human resources side of IT companies. In recent years, there has been a trend that is one of the cornerstones of digitalisation, namely the standard of digital education and the IT labour market. While a few years ago, both professional skills and salaries were below the Hungarian average, this seems to have reversed in the last 1-2 years. Based on market experience, average IT salaries in Romania are slowly catching up and, in some areas, even exceeding the salaries of Hungarian employees. A well-trained workforce can attract international capital in this knowledge-intensive industry. In the long term, this can greatly improve the country's competitiveness and prospects.

Hungary

Hungary's digital transformation is similar to that of Romania. The country has begun to catch up in the field of the digital economy and society in recent years, but it still lags behind the EU average. Following the change of regime, Hungary took initial steps to develop its information and communication technology infrastructure. The growth in internet access and mobile phone coverage has contributed significantly to laying the foundations for the digital economy. Its accession to the European Union in 2004 created new opportunities to accelerate digital transformation. EU funding has facilitated the development of digital skills, the introduction of e-government services and the growth of the IT sector. The 2010s were marked by digitalisation, which was given high priority in government programmes. Internet access expanded significantly, with residential internet usage rising from 55% in 2010 to 82% in 2023. (Éltető, 2021.) At the same time, an increasing proportion of the population used e-services: in 2010, 34% used the internet for administrative purposes, rising to 72% in 2023. The start-up ecosystem has also begun to develop , bringing innovative Hungarian companies such as Graphisoft, Prezi, LogMeIn, Ustream and NNG to the market. Currently, Hungary is below the EU average in terms of digital economy development. Based on the 2023 DESI index, Hungary ranks 21st among the 27 EU member states. The digital divide remains a significant problem, especially in rural areas and among lower-income groups. According to the 2023 Eurostat survey, 63% of the rural population and 44% of the lowest income quintile have internet access.

COVID has accelerated the digital transformation in almost all sectors, especially in the areas of digital infrastructure, remote working and education. Hungary has sought to respond to these challenges by supporting online education and work. From an economic perspective, however, it was the adaptability and flexibility of companies that helped them weather this period. Many companies had to rethink and reorganise their investments, especially in digital infrastructure. Even in larger multinational companies, it was not common for all employees to work with their own laptops in an office environment. Data centres and network infrastructure were not prepared for the fact that, from one day to the next, almost all employees would want to access company servers and applications remotely. By the end of the pandemic, it can generally be said that companies' digital maturity had improved thanks to the necessary investments.

Hungary has launched several digitalisation programmes and initiatives in recent years to promote the country's digital transformation, improve the competitiveness of companies and support the development of the digital skills of the population. These programmes are generally part of projects supported by the European Union, but they are also financed by Hungarian government funds. Some notable examples include:

Digital Hungary Strategy 2022-2030: The strategy aims to make Hungary a European leader in the digital economy by 2030. The five pillars of the strategy are: digital infrastructure, digital skills, digital government, digital economy and digital society. (kormany.hu, 2022)

Modern Enterprises Programme: The aim of the programme is to increase the competitiveness of small and medium-sized enterprises (SMEs) through the introduction of digital technologies. Under the programme, SMEs can receive advice, training and financing for digitalisation. (Chamber of Commerce and Industry, 2022)

Digital Wellbeing Programme: The aim of the programme is to develop the digital skills of the population and to promote the widespread use of digital tools. The programme provides free online training, digital education programmes and IT tools for the population. (Government IT Development Agency, 2023)

Central Electronic Administration System (KEÜR): An online platform that enables the population and businesses to conduct their business with the authorities electronically. Its aim is to create digital citizenship, establish a Hungarian-based cloud service, implement data-driven government decision-making and decision-making, and operate the coordinated infrastructure. (Digital Hungary Agency, 2022)

Hungary's 5G Strategy: The aim of the strategy is to make Hungary a European leader in 5G networks by 2025. The government will provide the necessary resources to implement the strategy and encourage the private sector to build 5G networks and develop 5G-based applications. The involvement of civil society ensures that the strategy also takes into account the needs of the population. (Government Information Technology Development Agency, 2023)

These are all forward-looking measures that will improve competitiveness in the long term. From a business perspective, the question is what these indicators will improve in relation to.

CONCLUSIONS

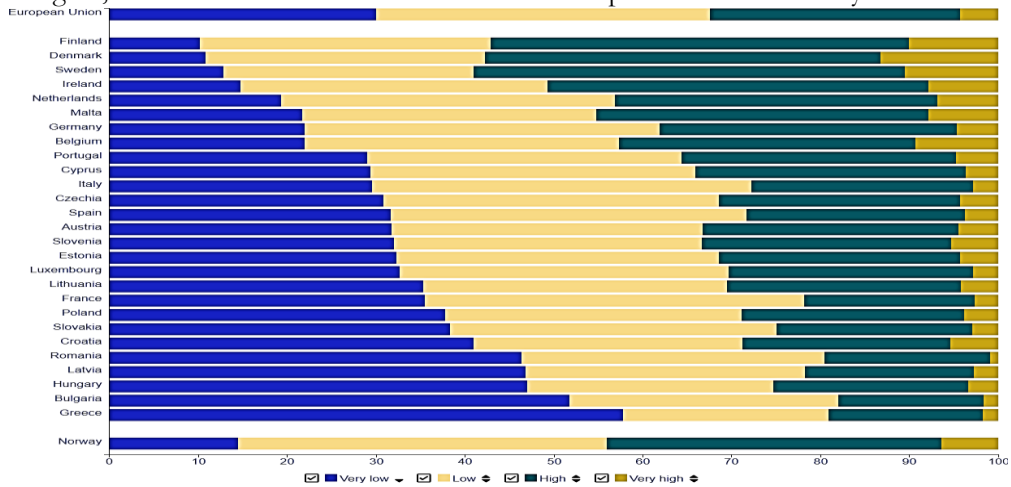
The purpose of data analysis is to draw forward-looking conclusions. There are many comparative statistics for the above countries, and they can be interpreted in many different ways. The following figure shows the "Digital Intensity Level of Enterprises", a composite indicator defined by Eurostat to measure the degree of digital technology adoption by enterprises in the European Union. It shows the extent to which an enterprise uses digital technologies in its operations and services. Eurostat uses several specific variables to determine the digital intensity level (DII) of a business. The most important of these are:

- Use of any artificial intelligence technology
- E-commerce sales accounting for at least 1% of total turnover
- Digital marketing, use of social media for marketing purposes
- Use of cloud-based computing services
- Digital communication, such as participation in online meetings

Based on the scores obtained from the combined factors, businesses can be classified into one of four DII levels:

- Very low: 0-3 points (in the figure: Very Low)
- Low: 4-6 points (in the figure: Low)
- High: 7-9 points (in the figure: High)
- Very high: 10-12 points (in the figure: Very High)

Estonia ranks in the middle, while Romania and Hungary are at the bottom. Only Bulgaria and Greece are behind us. In the figure, the blue colour indicates the number of companies that scored very low in this survey. This



does not necessarily mean that this figure is proportional to turnover. All statistics have weaknesses, which may

stem from the source of the data or the filtering of the data. Therefore, source criticism and contextualisation are particularly important.

Figure 21 Digital intensity level of businesses in 2022, X: countries, Y: distribution of DII levels % (EUROSTAT, 2022)

The distribution of enterprises can be observed in terms of how they perform in different categories. The 2022 data clearly shows the lagging performance of Hungarian enterprises, especially in the very low and low categories. This is why we are at the bottom of the field. However, if we look at the Very high category, we see that Hungary should be higher up the list. The advantage of a data-driven decision-making system is that we can dig deeper into the data and apply different filtering criteria to better understand the actual situation.

I performed a filter on the same data set, highlighting large companies. This refers to all companies with more than 250 employees or a turnover of more than EUR 50 million.

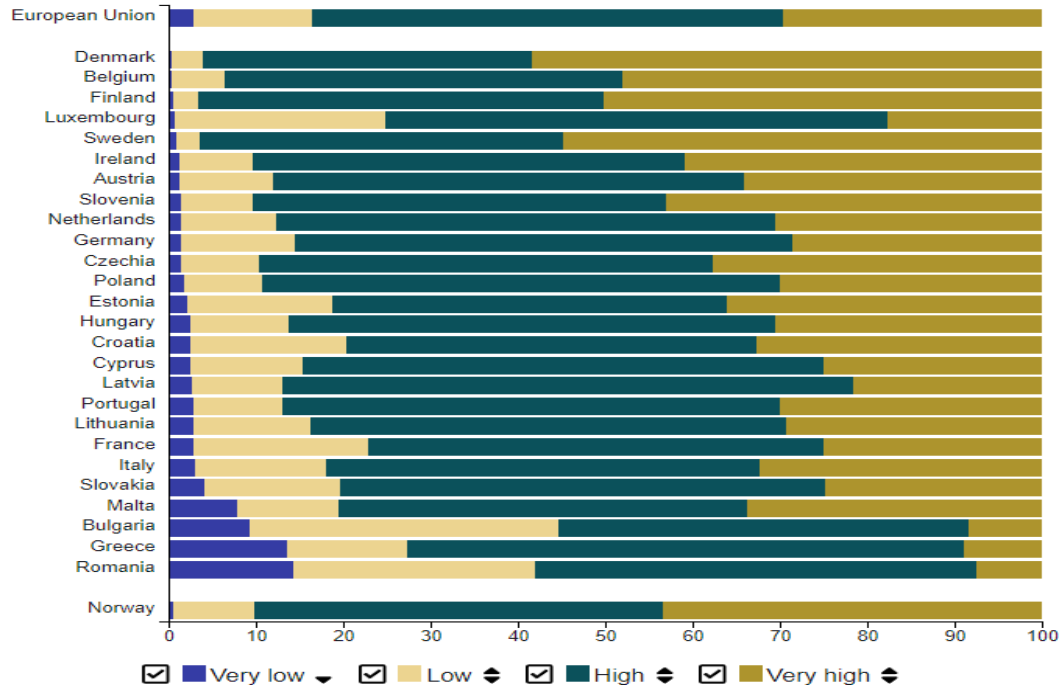


Figure 22 *Digital intensity level of large companies in 2022, X: countries, Y: DII level distribution %, Source: (EUROSTAT, 2022)*

Looking at this data series, Hungary ranks much higher. It is right in the middle of the pack, just behind Estonia. Romania ranks last. It is clear that the way data is presented can lead to different outcomes when making decisions. Expertise is essential for decision-making in corporate management, because two completely different conclusions can be drawn from the same data set if we do not know the context.

In this case, we can conclude that Hungary ranks reasonably well in terms of the digital culture of large companies. One reason for this is that multinational companies established subsidiaries in Hungary earlier and brought with them the frameworks and processes necessary for modern business intelligence. This has generated the expertise and experience that is indispensable in this field. However, it is also clear that this expertise has remained primarily with large companies, because small and medium-sized enterprises cannot afford to pay these specialists and cannot necessarily create the financial framework for the investments required for such systems. A well-functioning business intelligence system not only has set-up costs, but also ongoing operating costs. As a company grows, it needs to continuously develop its decision support systems. It needs to learn how to manage its data assets. The case study clearly shows that the information extracted from the data can be used not only for abstract decision support, but can also be magnetised in a concrete way, which can reduce costs or, in this case, lead to better customer retention, resulting in more stable and predictable revenues.

One of the goals of digitalisation efforts should be to improve competencies in this area on a broad scale, which requires a high level of education where future employees can be introduced to modern IT systems at an early stage. This is a complex problem that needs to be addressed at all levels of society.

RECOMMENDATIONS

One of the goals of my research is to gain insight into the relationship between data assets, digitalisation and competitiveness, both at the corporate and national economic levels. The experience I have gained in my work provides insight into some specific segments of this field. The highest possible level of digitisation represents a clear competitive advantage for companies and, thus, for the national economy as a whole. How can we achieve long-term sustainable development and maintain our competitiveness? I will attempt to make a few suggestions from a company's perspective.

Educational Cooperation

Information technology is one of the most dynamically developing industries, and keeping up with it requires serious effort. The traditional education system lacks the flexibility and responsiveness to always teach students the latest technologies. It is in the interest of companies to have well-trained, well-qualified career starters entering the labour market. A framework needs to be created to enable companies to establish closer cooperation with educational institutions, whether secondary schools or higher education institutions. There are already examples of this, but they are often ad hoc. In most cases, larger companies have the necessary expertise, but knowledge transfer and education require other skills. A framework in which all parties benefit from mutual cooperation could be effective. In many cases, senior employees who are proficient in a particular technology are unable to pass on their knowledge because they lack the necessary teaching skills. As a result, they do not pass on their knowledge adequately or in a structured manner, either within the company or to external students. Educational institutions could hold classes on teaching methodology and educational development for company experts, who could then give high-quality lectures at the educational institution or even teach entire professional subjects. This would bring in the latest technologies and trends and enable them to pass on marketable practical knowledge to students. Professionals and companies would also benefit, as the quality of their internal training would improve significantly. During the knowledge transfer, students would be able to get to know real companies and professionals and gain professional experience. Companies would gain insight into the graduate market and their professional competencies, and could later hire these students as employees or interns. This symbiosis would also increase the level of general digital competencies in the long term. Since educational institutions are mostly state organisations, it is important to have a transparent framework in which participants could be encouraged to engage in long-term cooperation, for example through various forms of support .

Corporate Synergies

For companies, especially small and medium-sized enterprises, the costs of digital innovation can place a serious burden on their budgets. Wider dissemination of digital innovation tenders can help to bridge this gap, but first and foremost, companies need to recognise the potential of proper use of data assets. However, in many cases, they lack the expertise to understand the direction in which they should develop. One solution here could be for larger IT consulting firms to act as business angels, helping small and medium-sized enterprises not with money but with expertise. In such cases, they would provide assistance in the form of surveys and advice, which in most cases only requires human resources. Such a tender framework would provide an opportunity to connect experienced companies with companies in need of development. In many cases, larger companies have temporary spare capacity that they are unable to utilise. In such cases, even a few hours of consulting by a business analyst or an expert in digital architecture could help a small business that would not be able to afford this at market prices. This can also be useful for the consulting firm, as it can utilise its temporary spare capacity and its people can gain experience by solving new problems and learning about new systems.

Company-Specific Research and Development

In large companies, it is often the case that development is continuous, but there is no capacity left for real research and development. Everyone is trying to keep up with the market and implement existing, tried-and-tested technologies, but there is no capacity left to carry out actual customised technological innovation. Above a certain company size, especially for those working in a technological sector or who have already achieved an acceptable level of digitalisation, it is worth setting up an innovation team. The main task of such a team would not be to solve current problems, but to search for technologies that are new or not yet widely used. So-called "Proof of Concept" projects provide an opportunity to test new innovative technologies on a small scale in a digital "laboratory" environment. It is possible to experiment with problems that cannot yet be fully implemented due to the current size or development of the company, or for which the framework does not yet exist, but which may be necessary in the future. A great deal can be learned during the implementation of such a small concept project, and when it comes to wider implementation later on, development can begin with actual experience. Such a project could be the analysis of part of the internal data assets or the testing of a technology among a small group of users,

such as Generative AI, which is now appearing everywhere, and other productivity-enhancing artificial intelligence applications. Such small-scale testing can highlight problems or prove the legitimacy of a wider rollout. Of course, all such projects have a cost, especially in terms of human resources. However, those who always just follow the market will not be able to gain a competitive advantage from the opportunities offered by new technologies.

SUMMARY

My research, entitled "Is data really the new oil?", analyses one of the most important trends in the modern business world, the significance of data-driven decision-making and its application, using the telecommunications sector as an example. How can valuable insights and deeper understanding be gained from the available data to facilitate strategic decision-making and improve corporate performance?

I presented the concept and significance of data assets, as well as the basics of data-driven decision-making. I emphasised the importance of treating data as a strategic resource in terms of competitiveness and corporate success. This was followed by a presentation of data analysis and methodologies, machine learning and the CRISP-DM framework, which describes the basic tools and processes of data science.

My primary research focused on a case study that concentrated on predicting subscriber churn, i.e. switching service providers. During the research, I applied and presented various modelling techniques – logistic regression and random forest algorithms – to show how machine learning models can help predict churn, thus supporting companies in developing customer retention strategies.

I examined the correlations of competitiveness by comparing countries to see what role digitalisation plays in the competitiveness of national economies. I paid particular attention to analysing the situation in Hungary, highlighting the relationship between digitalisation processes and competitiveness.

In my research, I examined three countries from an international perspective based on the European Commission's DESI (Digital Economy and Society Index) data: Hungary, Estonia and Romania, in terms of the integration of digital technologies. Estonia stands out in the field of digitalisation, occupying a leading position in this competition, while Hungary's advantage is slowly eroding compared to Romania.

Estonia, known as the "digital tiger", has achieved significant success in the digital economy thanks to early investments and the active involvement of its citizens. Government initiatives such as the "Tiger Leap" programme have focused on widespread digital education in schools and among the population, laying the foundation for the country's high internet usage rate and the spread of e-services. In the case of Hungary and Romania, the digital transformation is ongoing, although both countries lag behind the EU average. Both countries have taken significant steps to develop their information and communication technology infrastructure, which has contributed to laying the foundations for the digital economy and developing digital skills. EU funding and government programmes have facilitated digital transformation, particularly in the areas of e-government and e-health services. Overall, the international overview and country comparisons highlight the complex relationship between digitalisation and economic development. Although digital progress can offer advantages in terms of competitiveness and economic growth, the relationship is not linear, and other factors also influence countries' economic performance.

Drawing on my own experience, I have tried to make suggestions on how to raise digital literacy, how to build on existing competencies, and how to take IT education to a higher level through various collaborations, both within educational institutions and in the corporate environment. How to make better use of free capacity and help each other move forward. For larger companies, research and development and smaller pilot projects can help them capitalise on their data assets.

Data-driven decision-making and maximising the potential of data assets can provide a clear competitive advantage for companies that have the expertise to evaluate the results. It is in the interest of companies to invest in digitalisation, and it is also in the interest of national economies. The question is no longer whether investing in digitalisation will give us an advantage, but whether not investing will cause us to fall behind. Based on my research, the answer is clear to me.

REFERENCES

Bughin, J. (2017). *Digital America: A tale of the haves and have-mores.* McKinsey Global Institute. Coenen, F. (2021). Data mining: past, present and future. *Cambridge University Press*, 25 - 29.

Datacamp. (2 November 2023). *Data Science in Marketing: Customer Churn Rate Prediction*. Downloaded on 11 November 2023, source: Datacamp.com: https://www.datacamp.com/blog/data- science-in-marketing-customer-churn-rate-prediction?utm_source=google&utm_medium=paid_search&utm_campaignid=195897208

18&utm_adgroupid=152984009694&utm_device=c&utm_keyword=&utm_matchtype=&ut
m_network=g&utm_adpostion=&utm
Davenport, T. H. (2014). *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities.*
Harvard Business Review Press.
Diamantaras, K. (2 November 2020). *customer-churn-prediction-2020*. Downloaded on 21 September 2023, source: Kaggle.com: https://kaggle.com/competitions/customer-churn-prediction-2020
Digital Hungary Agency. (2022).*www.dmu.gov.hu/* . Downloaded on 17 October 2023, source: bemutatkozas-dmu: https://www.dmu.gov.hu/cikkek/bemutatkozas-dmu
Éltető, A. (2021). Digitalisation and location advantages. *Foreign Economy*, (pp. 91–105).
ERIK, T. M. (2011). Algorithmic trading: industry trend or market bubble?
CREDIT INSTITUTION REVIEW, TENTH VOLUME, *ISSUE 3*, 187.
European Commission. (27 September 2023). *Digital Strategy Europe*. Downloaded on 10 January 2024, source: https://digital-strategy.ec.europa.eu/: https://digital- strategy.ec.europa.eu/hu/policies/desi-hungary
EUROSTAT. (2022). *https://ec.europa.eu/eurostat*. Downloaded on 16 January 2024, source: digitalisation-2023: https://ec.europa.eu/eurostat/web/interactive- publications/digitalisation-2023
Eurostat. (2023). *https://ec.europa.eu/eurostat*. Downloaded on: 16 January 2024, source: Main GDP aggregates:
https://ec.europa.eu/eurostat/databrowser/view/NAMQ_10_GDP custom_7680558/book mark/table?lang=en&bookmarkId=a4ce6a9d-7ef1-48f1-a5bf-e23a717fcf75
Forbes. (2 March 2022). *Data as The New Oil Is Not Enough: Four Principles For Avoiding Data Fires*.
Downloaded on 15 October 2023, source: Forbes: https://www.forbes.com/sites/nishatalagala/2022/03/02/data-as-the-new-oil-is-not- enough-four-principles-for-avoiding-data-fires/?sh=1fa8c3a9c208
Glimsdahl, N. (17 August 2021). *https://www.icmi.com/* . Download date: 5 March 2024, source: ICMI: https://www.icmi.com/resources/2021/cx-impact-on-customer-churn
Glimsdahl, N. (11 September 2021). *ICMI*. Source: https://www.icmi.com/: https://www.icmi.com/resources/2021/cx-impact-on-customer-churn
IBM. (2016). *Analytics Solutions Unified Method - Implementations with Agile principles.* IBM. IBT. (15 May 2023). *The birth and evolution of artificial intelligence*. Source:
https://ibtconsulting.hu/: https://ibtconsulting.hu/blog/mesterseges-intelligencia-
szuletese-es-evolucioja
James, G. W. (2023). *An introduction to statistical learning: With applications in Python.* Springer.
Chamber of Commerce and Industry. (2022). *https://digitalisfogalomtar.vallalkozzdigitalisan.hu/*. Downloaded on 11 January 2024, source: MODERN BUSINESS PROGRAMME:
https://digitalisfogalomtar.vallalkozzdigitalisan.hu/
kormany.hu. (2022). *https://kormany.hu*. Downloaded on: 18 January 2024, source: national-digitalisation-strategy-2022-2030: https://kormany.hu/dokumentumtar/nemzeti- digitalisation-strategy-2022-2030
Government Information Technology Development Agency. (2023). *https://digitalisjoletprogram.hu/*. Downloaded on 11 January 2024, source: https://digitalisjoletprogram.hu/: https://digitalisjoletprogram.hu/
Luenendonk, M. (24 August 2022). *Exploring the Top 10 Data Types You Must Know*. Downloaded on 1 November 2023, source: FounderJar: https://www.founderjar.com/data-types/
McKinsey & Company. (5 November 2023). *Unlocking Value from Data: A Framework for Action*. Downloaded on 12 January 2024, source: https://www.mckinsey.com/: https://www.mckinsey.com/capabilities/quantumblack/our-insights/how-to-unlock-the- full-value-of-data-manage-it-like-a-product
Nelson, D. (23 September 2020). *Structured vs unstructured data*. Downloaded on 24 February 2024, source: https://www.unite.ai/: https://www.unite.ai/hu/struktur%C3%A1lt-vs- struktur%C3%A1latlan-adatok/
Novack, G. (5 June 2020). *towardsdatascience.com Building a One Hot Encoding Layer with TensorFlow*. Download date: 23 February 2024, source: towardsdatascience.com: https://towardsdatascience.com/building-a-one-hot-encoding-layer-with-tensorflow- f907d686bf39
Orlando Troisi, G. M. (2019). Growth hacking: Insights on data-driven decision-making from three firms. *Industrial Marketing Management*, 538-557.
Pichai, S. (3 April 2023). *Google DeepMind: Bringing together two world-class AI teams*. Downloaded on 22 February 2024, source: https://blog.google/: https://blog.google/technology/ai/april- ai-update/
Provost, F. &. (2013). *Data Science for Business.* O'Reilly Media, Inc.
SentinelOne Blog. (16 January 2023). *Top Cyber Attacks of 2023 (So Far).* Downloaded on 8 January 2024, source: SentinelOne Blog: https://www.sentinelone.com/blog/endpoint-identity-and-cloud- top-cyber-attacks-of-2023-so-far/
Stanford. (2021). The Turing Test. *Stanford Encyclopedia of Philosophy*.
Szentmihályi, S. (2023). Digitalisation and convergence – the example of Estonia. *Hitelintézeti Szemle*, 145–160.

Vyshnevskyi, O. (2021). Economic Growth In The Conditions Of Digitalisation In The EU Countries. *Studies of Applied Economics*.

Wikipedia. (10 November 2023). *Accuracy and precision*. Downloaded on 10 November 2023, source: Wikipedia: https://hu.wikipedia.org/wiki/Pontoss%C3%A1g_%C3%A9s_precizit%C3%A1s

